CONVERGENCE OF SPLIT-COMPLEX BACKPROPAGATION ALGORITHM WITH A MOMENTUM*

Huisheng Zhang†, *Wei Wu*‡

Abstract: This paper investigates a split-complex backpropagation algorithm with momentum (SCBPM) for complex-valued neural networks. Convergence results for SCBPM are proved under relaxed conditions and compared with the existing results. Monotonicity of the error function during the training iteration process is also guaranteed. Two numerical examples are given to support the theoretical findings.

Key words: *Complex-valued neural networks, split-complex backpropagation algorithm, convergence*

Received: November 18, 2008
Revised and accepted: January 10, 2011

1. Introduction

In recent years great interest has been aroused in the complex-valued neural networks (CVNNs) for their powerful capability in processing complex-valued signals [1, 2, 3]. CVNNs are extensions of real-valued neural networks [4, 5]. A fully complex backpropagation (BP) algorithm and a split-complex BP algorithm are two types of complex backpropagation algorithms for training CVNNs. Unlike the case of the fully complex BP algorithm [6, 7], the operation of activation function in the split-complex BP algorithm is split into a real part and an imaginary part [2, 4, 5, 8]. This split-complex BP algorithm avoids occurrence of singular points in the adaptive training process. This paper considers the split-complex BP algorithms.

*This work is partially supported by the National Natural Science Foundation of China (10871220,70971014), and by the Fundamental Research Funds for the Central Universities (2009QN075).

†Huisheng Zhang
Department of Mathematics, Dalian Maritime University, Dalian, China, 116026
‡Wei Wu – Corresponding author
School of Mathematical Sciences, Dalian University of Technology, Dalian, China, 116024, E-mail: wuweiw@dlut.edu.cn

Convergence properties of the algorithm are an important issue of its successful training for neural networks. The convergence of BP algorithm for real-valued neural networks has been analyzed by many authors from different aspects [9, 10, 11]. By using the contraction mapping theorem, the convergences in the mean and in the mean square for recurrent neurons were obtained by Mandic and Goh [12, 13]. For recent convergence analyses of complex-valued perceptrons and the BP algorithm for complex-valued neural networks, we refer the readers to [1, 14, 15].

It is well known that the learning process of BP algorithm can be very slow. To speed up and stabilize the learning procedure, a momentum is often added to the BP algorithm. The BP algorithm with a momentum (BPM in short) can be viewed as a memory gradient method in optimization theory [16]. Starting from a point close enough to a minimum point of the objective function, the memory gradient method converges under certain conditions [16, 17]. This local convergence result is also obtained in [18] where the convergence of BPM for a neural network is considered. However, these results cannot be applied to a more complex case where the initial weights are chosen stochastically. Some convergence results for the BPM are given in [19]; these results are of global nature since they are valid for arbitrarily given initial values of the weights. Our contribution in this paper is to present some convergence results of the split-complex backpropagation algorithm with a momentum (SCBPM). We borrow some ideas from [19], but we employ different proof techniques resulting in a new learning rate restriction which is much relaxed and easier to check than the counterpart in [19]. Actually, our approach can be applied to the BPM in [19] to relax the learning rate restriction there. Two numerical examples are given to support our theoretical findings. We also mention a recent paper on split quaternion nonlinear adaptive filtering [20], and we expect to extend our results to quaternion networks in future.

The rest of this paper is organized as follows. A CVNN model with one hidden layer and the SCBPM are described in the next section. Section 3 presents the main convergence theorem. Section 4 gives two numerical examples to support our theoretical findings. The proof of the theorem is presented in Section 5.

2. The Structure of the Network and the Learning Method

Fig. 1 shows the CVNN structure considered in this paper. It is a network with one hidden layer and one output neuron. Let the numbers of input and hidden neurons be L and M, respectively. We use "$(\cdot)^{\mathbf{T}}$" as the vector transpose operator, and write $\mathbf{w}_m = \mathbf{w}_m^R + i\mathbf{w}_m^I = (w_{m1}, w_{m2}, \cdots, w_{mL})^{\mathbf{T}} \in \mathbb{C}^L$ as the weight vector between the input neurons and m-th hidden neuron, where $w_{ml} = w_{ml}^R + iw_{ml}^I$, w_{ml}^R and $w_{ml}^I \in \mathbb{R}$, $i = \sqrt{-1}$, $m = 1, \cdots, M$, and $l = 1, \cdots, L$. Similarly, write $\mathbf{w}_{M+1} = \mathbf{w}_{M+1}^R + i\mathbf{w}_{M+1}^I = (w_{M+1,1}, w_{M+1,2}, \cdots, w_{M+1,M})^{\mathbf{T}} \in \mathbb{C}^M$ as the weight vector between the hidden neurons and the output neuron, where $w_{M+1,m} = w_{M+1,m}^R + iw_{M+1,m}^I$, $m = 1, \cdots, M$. For simplicity, all the weight vectors are incorporated into a total weight vector

$$\mathbf{W} = ((\mathbf{w}_1)^{\mathbf{T}}, (\mathbf{w}_2)^{\mathbf{T}}, \cdots, (\mathbf{w}_{M+1})^{\mathbf{T}})^{\mathbf{T}} \in \mathbb{C}^{ML+M}. \quad (1)$$

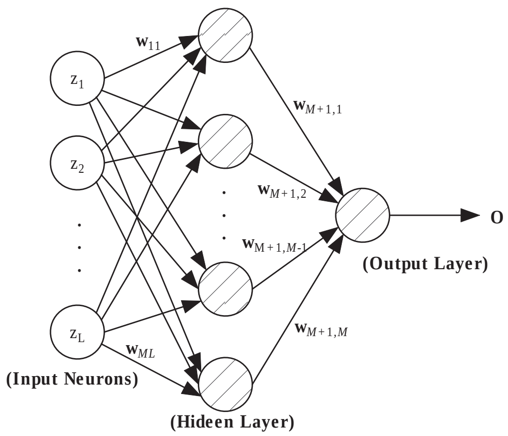

Fig. 1 *Structure of a CVNN.*

For an input signal $\mathbf{z} = (z_1, z_2, \cdots, z_L)^{\mathbf{T}} = \mathbf{x}+i\mathbf{y} \in \mathbb{C}^L$, where $\mathbf{x} = (x_1, x_2, \cdots, x_L)^{\mathbf{T}} \subset \mathbb{R}^L$, and $\mathbf{y} = (y_1, y_2, \cdots, y_L)^{\mathbf{T}} \subset \mathbb{R}^L$, the input to the m-th hidden neuron is

$$\begin{aligned} U_m &= U_m^R + iU_m^I \\ &= \sum_{l=1}^{L}(w_{ml}^R x_l - w_{ml}^I y_l) + i\sum_{l=1}^{L}(w_{ml}^I x_l + w_{ml}^R y_l) \\ &= \mathbf{x}^{\mathbf{T}}\mathbf{w}_m^R - \mathbf{y}^{\mathbf{T}}\mathbf{w}_m^I + i(\mathbf{x}^{\mathbf{T}}\mathbf{w}_m^I + \mathbf{y}^{\mathbf{T}}\mathbf{w}_m^R). \end{aligned} \tag{2}$$

In order to apply the SCBPM to train a CVNN, we consider the following popular real-imaginary-type activation function [22]

$$f_C(U) = f(U^R) + if(U^I) \tag{3}$$

for any $U = U^R + iU^I \in \mathbb{C}$, where f is a given real function (e.g., a sigmoid function). The output H_m of the hidden neuron m is given by:

$$H_m = H_m^R + iH_m^I = f(U_m^R) + if(U_m^I). \tag{4}$$

Similarly, the input to the output neuron is

$$\begin{aligned} S &= S^R + iS^I \\ &= \sum_{m=1}^{M}(w_{M+1,m}^R H_m^R - w_{M+1,m}^I H_m^I) + i\sum_{m=1}^{M}(w_{M+1,m}^I H_m^R + w_{M+1,m}^R H_m^I) \\ &= (\mathbf{H}^R)^{\mathbf{T}}\mathbf{w}_{M+1}^R - (\mathbf{H}^I)^{\mathbf{T}}\mathbf{w}_{M+1}^I + i\left((\mathbf{H}^R)^{\mathbf{T}}\mathbf{w}_{M+1}^I + (\mathbf{H}^I)^{\mathbf{T}}\mathbf{w}_{M+1}^R\right) \end{aligned} \tag{5}$$

and the output of the network is given by

$$O = O^R + iO^I = g(S^R) + ig(S^I), \tag{6}$$

where $\mathbf{H}^R = (H_1^R, H_2^R, \cdots, H_M^R)^\mathbf{T}, \mathbf{H}^I = (H_1^I, H_2^I, \cdots, H_M^I)^\mathbf{T}$, and g is a given real function. A bias weight can be added to each neuron, but it is omitted for simplicity of the presentation and deduction.

Let the network be supplied with a given set of training examples $\{\mathbf{z}^j, d^j\}_{j=1}^J \subset \mathbb{C}^L \times \mathbb{C}$. For each input $\mathbf{z}^j = \mathbf{x}^j + i\mathbf{y}^j$ $(1 \leq j \leq J)$ from the training set, we write $U_m^j = U_m^{j,R} + iU_m^{j,I}$ $(1 \leq m \leq M)$ as the input for the hidden neuron m, $H_m^j = H_m^{j,R} + iH_m^{j,I}$ $(1 \leq m \leq M)$ as the output for the hidden neuron m, $S^j = S^{j,R} + iS^{j,I}$ as the input to the output neuron, and $O^j = O^{j,R} + iO^{j,I}$ as the final output. The square error function of CVNN trained by the SCBPM can be represented as follows:

$$\begin{aligned} E(\mathbf{W}) &= \frac{1}{2}\sum_{j=1}^J (O^j - d^j)(O^j - d^j)^* \\ &= \frac{1}{2}\sum_{j=1}^J [(O^{j,R} - d^{j,R})^2 + (O^{j,I} - d^{j,I})^2] \\ &= \sum_{j=1}^J \left[g_{jR}(S^{j,R}) + g_{jI}(S^{j,I})\right], \end{aligned} \tag{7}$$

where "$(\cdot)^*$" denotes the complex conjugate operator, $d^{j,R}$ and $d^{j,I}$ are the real part and imaginary part of the desired output d^j respectively, and

$$g_{jR}(t) = \frac{1}{2}(g(t) - d^{j,R})^2, g_{jI}(t) = \frac{1}{2}(g(t) - d^{j,I})^2, t \in \mathbb{R}, 1 \leq j \leq J. \tag{8}$$

The purpose of the network training is to find $\mathbf{W}^\star$ to minimize $E(\mathbf{W})$. By writing

$$\begin{aligned} \mathbf{H}^j &= \mathbf{H}^{j,R} + i\mathbf{H}^{j,I} \\ &= (H_1^{j,R}, H_2^{j,R}, \cdots, H_M^{j,R})^\mathbf{T} + i(H_1^{j,I}, H_2^{j,I}, \cdots, H_M^{j,I})^\mathbf{T}, \end{aligned} \tag{9}$$

we get the gradients of $E(\mathbf{W})$ with respect to $\mathbf{w}_m^R$ and $\mathbf{w}_m^I$ respectively as follows

$$\frac{\partial E(\mathbf{W})}{\partial \mathbf{w}_{M+1}^R} = \sum_{j=1}^J \left[g'_{jR}(S^{j,R})\mathbf{H}^{j,R} + g'_{jI}(S^{j,I})\mathbf{H}^{j,I}\right], \tag{10}$$

$$\frac{\partial E(\mathbf{W})}{\partial \mathbf{w}_{M+1}^I} = \sum_{j=1}^J \left[-g'_{jR}(S^{j,R})\mathbf{H}^{j,I} + g'_{jI}(S^{j,I})\mathbf{H}^{j,R}\right], \tag{11}$$

$$\begin{aligned} \frac{\partial E(\mathbf{W})}{\partial \mathbf{w}_m^R} = \sum_{j=1}^J \Big[& g'_{jR}(S^{j,R})\big(w_{M+1,m}^R f'(U_m^{j,R})\mathbf{x}^j - w_{M+1,m}^I f'(U_m^{j,I})\mathbf{y}^j\big) \\ & + g'_{jI}(S^{j,I})\big(w_{M+1,m}^I f'(U_m^{j,R})\mathbf{x}^j + w_{M+1,m}^R f'(U_m^{j,I})\mathbf{y}^j\big)\Big], \quad 1 \leq m \leq M, \end{aligned} \tag{12}$$

$$\frac{\partial E(\mathbf{W})}{\partial \mathbf{w}_m^I} = \sum_{j=1}^J \Big[g'_{jR}(S^{j,R})\big(- w_{M+1,m}^R f'(U_m^{j,R})\mathbf{y}^j - w_{M+1,m}^I f'(U_m^{j,I})\mathbf{x}^j\big)$$

$$+ g'_{jI}(S^{j,I})\big(- w^I_{M+1,m} f'(U^{j,R}_m)\mathbf{y}^j + w^R_{M+1,m} f'(U^{j,I}_m)\mathbf{x}^j\big)\Big], \quad 1 \le m \le M. \tag{13}$$

Write $\mathbf{W}^n = ((\mathbf{w}^n_1)^\mathbf{T}, (\mathbf{w}^n_2)^\mathbf{T}, \cdots, (\mathbf{w}^n_{M+1})^\mathbf{T})^\mathbf{T}$ $(n = 0, 1, \cdots)$. Let $\mathbf{w}^0_m = \mathbf{w}^{0,R}_m + i\mathbf{w}^{0,I}_m$ be arbitrarily chosen initial weights. Let $\Delta\mathbf{w}^{0,R}_m = \Delta\mathbf{w}^{0,I}_m = 0$. Then the SCBPM algorithm updates the real part $\mathbf{w}^R_m$ and the imaginary part $\mathbf{w}^I_m$ of the weights $\mathbf{w}_m$ separately:

$$\begin{aligned} \Delta\mathbf{w}^{n+1,R}_m &= \mathbf{w}^{n+1,R}_m - \mathbf{w}^{n,R}_m = -\eta\frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}^R_m} + \tau^{n,R}_m \Delta\mathbf{w}^{n,R}_m, \\ \Delta\mathbf{w}^{n+1,I}_m &= \mathbf{w}^{n+1,I}_m - \mathbf{w}^{n,I}_m = -\eta\frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}^I_m} + \tau^{n,I}_m \Delta\mathbf{w}^{n,I}_m, \end{aligned} \tag{14}$$

where $\eta \in (0, 1)$ is the learning rate, $\tau^{n,R}_m$ and $\tau^{n,I}_m$ are the momentum factors, $m = 1, 2, \cdots, M + 1$, and $n = 0, 1, \cdots$.

For simplicity, let us denote

$$\begin{aligned} \mathbf{p}^{n,R}_m &= \frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}^R_m}, \\ \mathbf{p}^{n,I}_m &= \frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}^I_m}. \end{aligned} \tag{15}$$

Then (14) can be rewritten as

$$\begin{aligned} \Delta\mathbf{w}^{n+1,R}_m &= \tau^{n,R}_m \Delta\mathbf{w}^{n,R}_m - \eta\mathbf{p}^{n,R}_m, \\ \Delta\mathbf{w}^{n+1,I}_m &= \tau^{n,I}_m \Delta\mathbf{w}^{n,I}_m - \eta\mathbf{p}^{n,I}_m. \end{aligned} \tag{16}$$

Similarly to the BPM [19], we choose the adaptive momentum factor $\tau^{n,R}_m$ and $\tau^{n,I}_m$ as follows

$$\tau^{n,R}_m = \begin{cases} \frac{\tau\|\mathbf{p}^{n,R}_m\|}{\|\Delta\mathbf{w}^{n,R}_m\|}, & \text{if } \|\Delta\mathbf{w}^{n,R}_m\| \neq 0, \\ 0, & \text{else}, \end{cases} \tag{17}$$

$$\tau^{n,I}_m = \begin{cases} \frac{\tau\|\mathbf{p}^{n,I}_m\|}{\|\Delta\mathbf{w}^{n,I}_m\|}, & \text{if } \|\Delta\mathbf{w}^{n,I}_m\| \neq 0, \\ 0, & \text{else}, \end{cases} \tag{18}$$

where $\tau \in (0, 1)$ is a constant parameter and $\|\cdot\|$ is the usual Euclidian norm.

3. Main Results

The following assumptions will be used in our discussion.

$(A1)$**:** There exists a constant $C_1 > 0$ such that

$$\max_{t\in\mathbb{R}}\{|f(t)|, |g(t)|, |f'(t)|, |g'(t)|, |f''(t)|, |g''(t)|\} \le C_1.$$

$(A2)$**:** There exists a constant $C_2 > 0$ such that $\|\mathbf{w}^{n,R}_{M+1}\| \le C_2$ and $\|\mathbf{w}^{n,I}_{M+1}\| \le C_2$ for all $n = 0, 1, 2, \cdots$.

$(A3)$**:** The set $\Phi_0 = \{\mathbf{W} | \frac{\partial E(\mathbf{W})}{\partial \mathbf{w}^R_m} = 0, \frac{\partial E(\mathbf{W})}{\partial \mathbf{w}^I_m} = 0, m = 1, \cdots, M+1\}$ contains only finite points.

Theorem 1 *Suppose that Assumptions* (*A*1) *and* (*A*2) *are valid, and that* $\{\mathbf{w}_m^n\}$ *are the weight vector sequences generated by* (14)*. Then, there exists a constant* $C^\star > 0$ *such that for* $0 < s < 1, \tau = s\eta$ *and* $\eta \le \frac{1-s}{C^\star(1+s)^2}$*, the following results hold:*

(*i*) $E(\mathbf{W}^{n+1}) \le E(\mathbf{W}^n), \quad n = 0, 1, 2, \cdots$*;*

(*ii*) *There is* $E^\star \ge 0$ *such that* $\lim_{n\to\infty} E(\mathbf{W}^n) = E^\star$*;*

(*iii*) $\lim_{n\to\infty} \left\| \frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}_m^R} \right\| = 0$ *and* $\lim_{n\to\infty} \left\| \frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}_m^I} \right\| = 0, \quad 0 \le m \le M+1.$

Furthermore, if Assumption (*A*3) *also holds, then there exists a point* $W^\star \in \Phi_0$ *such that*

(*iv*) $\lim_{n\to\infty} \mathbf{W}^n = \mathbf{W}^\star.$

The monotonicity and convergence of the error function $E(\mathbf{W})$ during the learning process are shown in Conclusions (i) and (ii), respectively. Conclusion (iii) indicates the convergence of $\frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}_m^R}$ and $\frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}_m^I}$, referred to as weak convergence. The strong convergence of $\mathbf{W}^n$ is given in Conclusion (iv). We note that the restriction $\eta \le \frac{1-s}{C^\star(1+s)^2}$ in Theorem 1 is less restrictive and easier to check than the corresponding condition in [19]. We also mention that our results are of deterministic nature compared with a related work in [1], where the convergences in the mean and in the mean square for complex-valued perceptrons are obtained.

4. Numerical Example

In the following subsections we illustrate the convergence behavior of the SCBPM by using two numerical examples. In the both examples, we set the transfer function to be *tansig*(·) in MATLAB, which is a commonly used sigmoid function, and carry out 10 independent tests with the initial components of the weights stochastically chosen in [-0.5, 0.5]. The average of errors and the average of gradient norms for all the tests in each example are plotted.

4.1 XOR problem

The well-known XOR problem is a benchmark used in literature on neural networks. As in [22], the training samples of the encoded XOR problem for a CVNN is presented as follows:

$$\{\mathbf{z}^1 = -1 - i, d^1 = 1\}, \{\mathbf{z}^2 = -1 + i, d^2 = 0\},$$
$$\{\mathbf{z}^3 = 1 - i, d^3 = 1 + i\}, \{\mathbf{z}^4 = 1 + i, d^4 = i\}.$$

This example uses a network with one input node, three hidden nodes, and one output node. The learning rate η and the momentum parameter τ are set to be 0.1 and 0.01, respectively. The simulation results are presented in Fig. 2, which shows that the gradient tends to zero and the square error decreases monotonically as the number of iteration increases and at last it tends to a constant. This supports our theoretical findings.

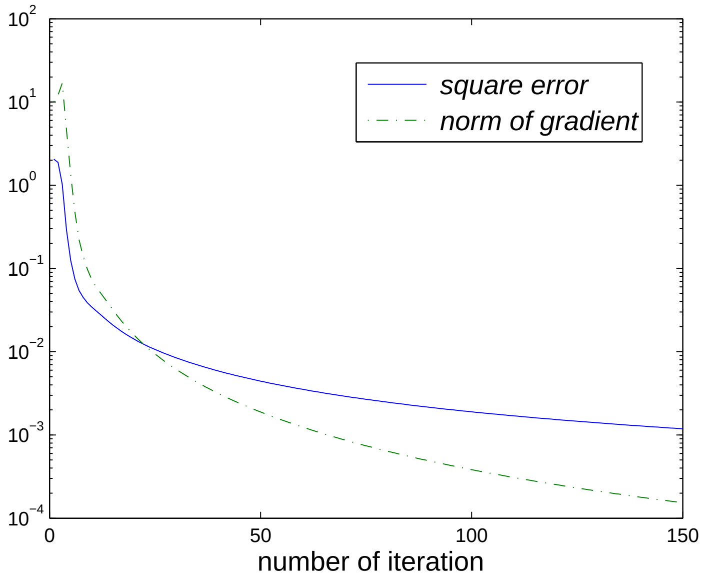

Fig. 2 *Convergence behavior of SCBPM for solving the XOR problem (norm of gradient* $= \sum_{m=1}^{M+1} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2)$*).*

4.2 Approximation problem

In this example, the synthetic complex-valued function [23] defined as

$$h(\mathbf{z}) = (z_1)^2 + (z_2)^2 \tag{19}$$

is approximated, where $\mathbf{z}$ is a two dimensional complex-valued vector comprised of z_1 and z_2. 10000 input points are selected from an evenly spaced $10 \times 10 \times 10 \times 10$ grid on $-0.5 \le Re(z_1), Im(z_1), Re(z_2), Im(z_2) \le 0.5$, where $Re(\cdot)$ and $Im(\cdot)$ represent the real part and imaginary part of a complex number, respectively. We use a network with 2 input neurons, 25 hidden neurons and 1 output neuron. The learning rate η and the momentum parameter τ are set to be 0.1 and 0.01, respectively. Fig. 3 illustrates the simulation results, which also support our convergence theorem.

5. Proofs

In this section, we first present two lemmas, then we use them to prove the main theorem.

Lemma 1 *Suppose that* $\mathcal{E} : R^{2ML+2M} \longrightarrow R$ *is continuous and differentiable on a compact set* $\Phi \subset R^{2ML+2M}$*, and that* $\Phi_1 = \{\mathbf{v} | \frac{\partial \mathcal{E}(\mathbf{v})}{\partial \mathbf{v}} = 0\}$ *contains only finite*

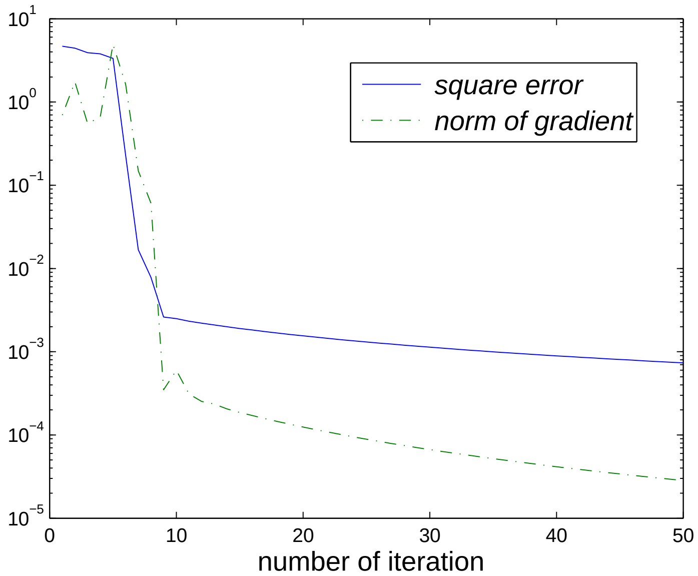

Fig. 3 *Convergence behavior of SCBPM for solving the approximation problem (norm of gradient* $= \sum_{m=1}^{M+1}(\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2)$*).*

points. If a sequence $\{\mathbf{v}^n\}_{n=1}^{\infty} \subset \Phi$ *satisfies*

$$\lim_{n\to\infty} \|\mathbf{v}^{n+1} - \mathbf{v}^n\| = 0, \ \lim_{n\to\infty} \|\nabla\mathcal{E}(\mathbf{v}^n)\| = 0,$$

then there exists a point $\mathbf{v}^\star \in \Phi_1$ *such that* $\lim_{n\to\infty} \mathbf{v}^n = \mathbf{v}^\star$.

Proof This result is almost the same as Theorem 14.1.5 in [21], and the detail of the proof is omitted. □

For any $1 \le j < J$, $1 \le m \le M$ and $n = 0, 1, 2, \cdots$, the following symbols will be used in our proof later on:

$$\begin{aligned}
&U_m^{n,j} = U_m^{n,j,R} + iU_m^{n,j,I} = (\mathbf{x}^j)^{\mathbf{T}}\mathbf{w}_m^{n,R} - (\mathbf{y}^j)^{\mathbf{T}}\mathbf{w}_m^{n,I} + i\left((\mathbf{x}^j)^{\mathbf{T}}\mathbf{w}_m^{n,I} + (\mathbf{y}^j)^{\mathbf{T}}\mathbf{w}_m^{n,R}\right),\\
&H_m^{n,j} = H_m^{n,j,R} + iH_m^{n,j,I} = f(U_m^{n,j,R}) + if(U_m^{n,j,I}),\\
&\mathbf{H}^{n,j,R} = (H_1^{n,j,R}, \cdots, H_M^{n,j,R})^{\mathbf{T}}, \mathbf{H}^{n,j,I} = (H_1^{n,j,I}, \cdots, H_M^{n,j,I})^{\mathbf{T}},\\
&S^{n,j} = S^{n,j,R} + iS^{n,j,I} = (\mathbf{H}^{n,j,R})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,R} - (\mathbf{H}^{n,j,I})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,I} + i\Big((\mathbf{H}^{n,j,R})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,I} +\\
&\qquad + (\mathbf{H}^{n,j,I})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,R}\Big),\\
&\boldsymbol{\psi}^{n,j,R} = \mathbf{H}^{n+1,j,R} - \mathbf{H}^{n,j,R}, \boldsymbol{\psi}^{n,j,I} = \mathbf{H}^{n+1,j,I} - \mathbf{H}^{n,j,I}.
\end{aligned} \tag{20}$$

Lemma 2 *Suppose Assumptions (A1) and (A2) hold, then for any* $1 \le j \le J$ *and* $n = 0, 1, 2, \cdots$, *we have*

$$|O^{j,R}| \le C_0, |O^{j,I}| \le C_0, \|\mathbf{H}^{n,j,R}\| \le C_0, \|\mathbf{H}^{n,j,I}\| \le C_0, \tag{21}$$

$$|g'_{jR}(t)| \le C_3, |g'_{jI}(t)| \le C_3, |g''_{jR}(t)| \le C_3, |g''_{jI}(t)| \le C_3, t \in \mathbb{R}, \tag{22}$$

$$\max\{\|\boldsymbol{\psi}^{n,j,R}\|^2, \|\boldsymbol{\psi}^{n,j,I}\|^2\} \le C_4(\tau+\eta)^2 \sum_{m=1}^{M} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2), \tag{23}$$

$$\begin{aligned}\sum_{j=1}^{J} \Big(g'_{jR}(S^{n,j,R}) \Big((\mathbf{H}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,R} - (\mathbf{H}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,I} \Big) \\ + g'_{jI}(S^{n,j,I}) \Big((\mathbf{H}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,I} + (\mathbf{H}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,R} \Big) \Big) \\ \le (\tau - \eta) \big(\|\mathbf{p}_{M+1}^{n,R}\|^2 + \|\mathbf{p}_{M+1}^{n,I}\|^2 \big),\end{aligned} \tag{24}$$

$$\begin{aligned}\sum_{j=1}^{J} \Big(g'_{jR}(S^{n,j,R}) \Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \mathbf{w}_{M+1}^{n,R} - (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \mathbf{w}_{M+1}^{n,I} \Big) \\ + g'_{jI}(S^{n,j,I}) \Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \mathbf{w}_{M+1}^{n,I} + (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \mathbf{w}_{M+1}^{n,R} \Big) \Big) \\ \le (\tau - \eta + C_5(\tau+\eta)^2) \sum_{m=1}^{M} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2),\end{aligned} \tag{25}$$

$$\begin{aligned}\sum_{j=1}^{J} \Big(g'_{jR}(S^{n,j,R}) \Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,R} - (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,I} \Big) \\ + g'_{jI}(S^{n,j,I}) \Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,I} + (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,R} \Big) \Big) \\ \le C_6(\tau+\eta)^2 \sum_{m=1}^{M+1} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2),\end{aligned} \tag{26}$$

$$\begin{aligned}\frac{1}{2} \sum_{j=1}^{J} (g''_{jR}(t_1^{n,j})(S^{n+1,j,R} - S^{n,j,R})^2 + g''_{jI}(t_2^{n,j})(S^{n+1,j,I} - S^{n,j,I})^2) \\ \le C_7(\tau+\eta)^2 \sum_{m=1}^{M+1} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2)),\end{aligned} \tag{27}$$

where C_i $(i = 0, 3, \cdots, 7)$ *are constants independent of* n *and* j, *each* $t_1^{n,j} \in \mathbb{R}$ *lies on the segment between* $S^{n+1,j,R}$ *and* $S^{n,j,R}$, *and each* $t_2^{n,j} \in \mathbb{R}$ *lies on the segment between* $S^{n+1,j,I}$ *and* $S^{n,j,I}$.

Proof The validation of (21) results easily from (4)–(6) when the set of samples is fixed and Assumptions $(A1)$ and $(A2)$ are satisfied. By (8), we have

$$\begin{aligned} g'_{jR}(t) &= g'(t)(g(t) - O^{j,R}), \\ g'_{jI}(t) &= g'(t)(g(t) - O^{j,I}), \\ g''_{jR}(t) &= g''(t)(g(t) - O^{j,R}) + (g'(t))^2, \end{aligned}$$

$$g''_{jI}(t) = g''(t)(g(t) - O^{j,I}) + (g'(t))^2, \;\; 1 \le j \le J, \; t \in \mathbb{R}.$$

Then (22) follows directly from Assumption $(A1)$ by defining $C_3 = C_1(C_1 + C_0) + (C_1)^2$.

By (16), (17) and (18), for $m = 1, \cdots, M+1$,

$$\|\Delta \mathbf{w}_m^{n+1,R}\| = \|\tau_m^{n,R} \Delta \mathbf{w}_m^{n,R} - \eta \mathbf{p}_m^{n,R}\| \le \tau_m^{n,R} \|\Delta \mathbf{w}_m^{n,R}\| + \eta \|\mathbf{p}_m^{n,R}\| \le (\tau + \eta)\|\mathbf{p}_m^{n,R}\|. \tag{28}$$

Similarly, we have

$$\|\Delta \mathbf{w}_m^{n+1,I}\| \le (\tau + \eta)\|\mathbf{p}_m^{n,I}\|. \tag{29}$$

If Assumption $(A1)$ is valid, it follows from (20), (28), (29), the Cauchy-Schwartz Inequality and the Mean-Value Theorem for multivariate functions that for any $1 \le j \le J$ and $n = 0, 1, 2, \cdots$

$$\begin{aligned}
\|\boldsymbol{\psi}^{n,j,R}\|^2 = \|\mathbf{H}^{n+1,j,R} - \mathbf{H}^{n,j,R}\|^2 &= \left\| \begin{pmatrix} f(U_1^{n+1,j,R}) - f(U_1^{n,j,R}) \\ \vdots \\ f(U_M^{n+1,j,R}) - f(U_M^{n,j,R}) \end{pmatrix} \right\|^2 \\
&= \left\| \begin{pmatrix} f'(s_1^{n,j})((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}_1^{n+1,R} - (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}_1^{n+1,I}) \\ \vdots \\ f'(s_M^{n,j})((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}_M^{n+1,R} - (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}_M^{n+1,I}) \end{pmatrix} \right\|^2 \\
&= \sum_{m=1}^{M} (f'(s_m^{n,j})((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}_m^{n+1,R} - (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}_m^{n+1,I}))^2 \\
&\le 2(C_1)^2 \sum_{m=1}^{M} (((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}_m^{n+1,R})^2 + ((\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}_m^{n+1,I})^2) \\
&\le 2(C_1)^2 \sum_{m=1}^{M} (\|\Delta \mathbf{w}_m^{n+1,R}\|^2 \|\mathbf{x}^j\|^2 + \|\Delta \mathbf{w}_m^{n+1,I}\|^2 \|\mathbf{y}^j\|^2) \\
&\le C_4 \sum_{m=1}^{M} (\|\Delta \mathbf{w}_m^{n+1,R}\|^2 + \|\Delta \mathbf{w}_m^{n+1,I}\|^2) \\
&\le C_4 (\tau + \eta)^2 \sum_{m=1}^{M} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2),
\end{aligned} \tag{30}$$

where $C_4 = 2(C_1)^2 \max_{1 \le j \le J} \{\|\mathbf{x}^j\|^2, \|\mathbf{y}^j\|^2\}$ and $s_m^{n,j}$ is on the segment between $U_m^{n+1,j,R}$ and $U_m^{n,j,R}$ for $m = 1, \cdots, M$. Similarly we can get

$$\|\boldsymbol{\psi}^{n,j,I}\| \le C_4 (\tau + \eta)^2 \sum_{m=1}^{M} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2). \tag{31}$$

Thus, we have (23).

According to (16), (17) and (18), we have

$$
\begin{aligned}
&(\mathbf{p}_m^{n,R})^{\mathbf{T}}\Delta\mathbf{w}_m^{n+1,R} + (\mathbf{p}_m^{n,I})^{\mathbf{T}}\Delta\mathbf{w}_m^{n+1,I} \\
&\quad = -\eta\big(\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2\big) + \tau_m^{n,R}(\mathbf{p}_m^{n,R})^{\mathbf{T}}\Delta\mathbf{w}_m^{n,R} + \tau_m^{n,I}(\mathbf{p}_m^{n,I})^{\mathbf{T}}\Delta\mathbf{w}_m^{n,I} \\
&\quad \le -\eta\big(\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2\big) + \tau_m^{n,R}\|\Delta\mathbf{w}_m^{n,R}\|\|\mathbf{p}_m^{n,R}\| + \tau_m^{n,I}\|\Delta\mathbf{w}_m^{n,I}\|\|\mathbf{p}_m^{n,I}\| \\
&\quad \le (\tau-\eta)\big(\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2\big), \qquad (32)
\end{aligned}
$$

where $m = 1, \cdots, M, M+1$. This together with (10) and (11) validates (24):

$$
\begin{aligned}
&\sum_{j=1}^{J}\Big(g'_{jR}(S^{n,j,R})\Big((\mathbf{H}^{n,j,R})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,R} - (\mathbf{H}^{n,j,I})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,I}\Big) \\
&\qquad\qquad + g'_{jI}(S^{n,j,I})\Big((\mathbf{H}^{n,j,R})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,I} + (\mathbf{H}^{n,j,I})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,R}\Big)\Big) \\
&= \sum_{j=1}^{J}\big(g'_{jR}(S^{n,j,R})(\mathbf{H}^{n,j,R})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,R} + g'_{jI}(S^{n,j,I})(\mathbf{H}^{n,j,I})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,R} \\
&\qquad - g'_{jR}(S^{n,j,R})(\mathbf{H}^{n,j,I})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,I} + g'_{jI}(S^{n,j,I})(\mathbf{H}^{n,j,R})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,I}\big) \\
&= (\mathbf{p}_{M+1}^{n,R})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,R} + (\mathbf{p}_{M+1}^{n,I})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,I} \\
&\le (\tau-\eta)\big(\|\mathbf{p}_{M+1}^{n,R}\|^2 + \|\mathbf{p}_{M+1}^{n,I}\|^2\big). \qquad (33)
\end{aligned}
$$

Next, we prove (25). By (2), (4), (20) and Taylor's formula, for any $1 \le j \le J$, $1 \le m \le M$ and $n = 0, 1, 2, \cdots$, we have

$$
\begin{aligned}
H_m^{n+1,j,R} - H_m^{n,j,R} &= f(U_m^{n+1,j,R}) - f(U_m^{n,j,R}) \\
&= f'(U_m^{n,j,R})(U_m^{n+1,j,R} - U_m^{n,j,R}) + \frac{1}{2}f''(t_m^{n,j,R})(U_m^{n+1,j,R} - U_m^{n,j,R})^2
\end{aligned} \qquad (34)
$$

and

$$
\begin{aligned}
H_m^{n+1,j,I} - H_m^{n,j,I} &= f(U_m^{n+1,j,I}) - f(U_m^{n,j,I}) \\
&= f'(U_m^{n,j,I})(U_m^{n+1,j,I} - U_m^{n,j,I}) + \frac{1}{2}f''(t_m^{n,j,I})(U_m^{n+1,j,I} - U_m^{n,j,I})^2, \qquad (35)
\end{aligned}
$$

where $t_m^{n,j,R}$ is an intermediate point on the line segment between the two points $U_m^{n+1,j,R}$ and $U_m^{n,j,R}$, and $t_m^{n,j,I}$ between the two points $U_m^{n+1,j,I}$ and $U_m^{n,j,I}$. Thus, according to (2), (12), (13), (14), (20), (32) and (34)–(35) we have

$$
\begin{aligned}
&\sum_{j=1}^{J}\Big(g'_{jR}(S^{n,j,R})\Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,R} - (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,I}\Big) \\
&\qquad\qquad + g'_{jI}(S^{n,j,I})\Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,I} + (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,R}\Big)\Big) \\
&= \sum_{j=1}^{J}\sum_{m=1}^{M}\Big(g'_{jR}(S^{n,j,R})w_{M+1,m}^{n,R}f'(U_m^{n,j,R})\big((\mathbf{x}^j)^{\mathbf{T}}\Delta\mathbf{w}_m^{n+1,R} - (\mathbf{y}^j)^{\mathbf{T}}\Delta\mathbf{w}_m^{n+1,I}\big) \\
&\qquad - g'_{jR}(S^{n,j,R})w_{M+1,m}^{n,I}f'(U_m^{n,j,I})\big((\mathbf{x}^j)^{\mathbf{T}}\Delta\mathbf{w}_m^{n+1,I} + (\mathbf{y}^j)^{\mathbf{T}}\Delta\mathbf{w}_m^{n+1,R}\big)
\end{aligned}
$$

$$+ g'_{jI}(S^{n,j,I}) w^{n,I}_{M+1,m} f'(U^{n,j,R}_m) \left((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_m - (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_m \right)$$
$$+ g'_{jI}(S^{n,j,I}) w^{n,R}_{M+1,m} f'(U^{n,j,I}_m) \left((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_m + (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_m \right) \Big) + \delta_1$$
$$= \sum_{m=1}^{M} \Bigg(\bigg(\sum_{j=1}^{J} \Big[g'_{jR}(S^{n,j,R}) \left(w^{n,R}_{M+1,m} f'(U^{n,j,R}_m) \mathbf{x}^j - w^{n,I}_{M+1,m} f'(U^{n,j,I}_m) \mathbf{y}^j \right)$$
$$+ g'_{jI}(S^{n,j,I}) \left(w^{n,I}_{M+1,m} f'(U^{n,j,R}_m) \mathbf{x}^j + w^{n,R}_{M+1,m} f'(U^{n,j,I}_m) \mathbf{y}^j \right) \Big] \bigg)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_m$$
$$+ \bigg(\sum_{j=1}^{J} \Big[g'_{jR}(S^{n,j,R}) \left(-w^{n,R}_{M+1,m} f'(U^{n,j,R}_m) \mathbf{y}^j - w^{n,I}_{M+1,m} f'(U^{n,j,I}_m) \mathbf{x}^j \right)$$
$$+ g'_{jI}(S^{n,j,I}) \left(-w^{n,I}_{M+1,m} f'(U^{n,j,R}_m) \mathbf{y}^j + w^{n,R}_{M+1,m} f'(U^{n,j,I}_m) \mathbf{x}^j \right) \Big] \bigg)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_m \Bigg) + \delta_1$$
$$= \sum_{m=1}^{M} \left((\mathbf{p}^{n,R}_m)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_m + (\mathbf{p}^{n,I}_m)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_m \right) + \delta_1$$
$$\leq (\tau - \eta) \sum_{m=1}^{M} \left(\|\mathbf{p}^{n,I}_m\|^2 + \|\mathbf{p}^{n,R}_m\|^2 \right) + \delta_1, \tag{36}$$

where

$$\delta_1 = \frac{1}{2} \sum_{j=1}^{J} \sum_{m=1}^{M} \Big(g'_{jR}(S^{n,j,R}) w^{n,R}_{M+1,m} f''(t^{n,j,R}_m) \left((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_m + (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_m \right)^2$$
$$- g'_{jR}(S^{n,j,R}) w^{n,I}_{M+1,m} f''(t^{n,j,I}_m) \left((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_m + (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_m \right)^2$$
$$+ g'_{jI}(S^{n,j,I}) w^{n,I}_{M+1,m} f''(t^{n,j,R}_m) \left((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_m + (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_m \right)^2$$
$$+ g'_{jI}(S^{n,j,I}) w^{n,R}_{M+1,m} f''(t^{n,j,I}_m) \left((\mathbf{x}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_m + (\mathbf{y}^j)^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_m \right)^2 \Big). \tag{37}$$

Using Assumptions $(A1)$ and $(A2)$, (17), (18), (22) and the triangular inequality we immediately get

$$\delta_1 \leq |\delta_1| \leq C_5 \sum_{m=1}^{M} (\|\Delta \mathbf{w}^{n+1,R}_m\|^2 + \|\Delta \mathbf{w}^{n+1,I}_m\|^2)$$
$$\leq C_5 \sum_{m=1}^{M} \left((\tau^{n,R}_m \|\Delta \mathbf{w}^{n,R}_m\| + \eta \|\mathbf{p}^{n,R}_m\|)^2 + (\tau^{n,I}_m \|\Delta \mathbf{w}^{n,I}_m\| + \eta \|\mathbf{p}^{n,I}_m\|)^2 \right)$$
$$\leq C_5 (\tau + \eta)^2 \sum_{m=1}^{M} (\|\mathbf{p}^{n,R}_m\|^2 + \|\mathbf{p}^{n,I}_m\|^2), \tag{38}$$

where $C_5 = 2JC_1C_2C_3 \max_{1 \leq j \leq J} \{\|x^j\|^2 + \|y^j\|^2\}$. Now, (25) results from (36) and (38).

According to (20), (22), (23) and (28) we have

$$\sum_{j=1}^{J} \Big(g'_{jR}(S^{n,j,R}) \left((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}^{n+1,R}_{M+1} - (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}^{n+1,I}_{M+1} \right)$$

$$
\begin{aligned}
&\quad + g'_{jI}(S^{n,j,I})\left((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,I} + (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,R}\right)\Big) \\
&\le C_3\sum_{j=1}^{J}\left(\|\Delta\mathbf{w}_{M+1}^{n+1,R}\|^2 + \|\Delta\mathbf{w}_{M+1}^{n+1,I}\|^2 + \|\boldsymbol{\psi}^{n,j,R}\|^2 + \|\boldsymbol{\psi}^{n,j,I}\|^2\right) \\
&\le C_3\sum_{j=1}^{J}\left((\tau+\eta)^2(\|\mathbf{p}_{M+1}^{n,R}\|^2 + \|\mathbf{p}_{M+1}^{n,I}\|^2) + 2C_4(\tau+\eta)^2\sum_{m=1}^{M}(\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2)\right) \\
&\le C_6(\tau+\eta)^2\sum_{m=1}^{M+1}(\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2) \qquad (39)
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{J}(g''_{jR}(t_1^{n,j})(S^{n+1,j,R} - S^{n,j,R})^2 + g''_{jI}(t_2^{n,j})(S^{n+1,j,I} - S^{n,j,I})^2) \\
&\le \frac{C_3}{2}\sum_{j=1}^{J}((S^{n+1,j,R} - S^{n,j,R})^2 + (S^{n+1,j,I} - S^{n,j,I})^2) \\
&\le \frac{C_3}{2}\sum_{j=1}^{J}\Bigg(\Big((\mathbf{H}^{n+1,j,R})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,R} - (\mathbf{H}^{n+1,j,I})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,I} + (\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,R} - \\
&\qquad -(\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,I}\Big)^2 \\
&\quad + \Big((\mathbf{H}^{n+1,j,R})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,I} + (\mathbf{H}^{n+1,j,I})^{\mathbf{T}}\Delta\mathbf{w}_{M+1}^{n+1,R} + (\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,I} + \\
&\qquad +(\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}}\mathbf{w}_{M+1}^{n,R}\Big)^2\Bigg) \\
&\le 2C_3J\max\{(C_0)^2 + (C_2)^2\}(\|\Delta\mathbf{w}_{M+1}^{n+1,R}\|^2 + \|\Delta\mathbf{w}_{M+1}^{n+1,I}\|^2 + \|\boldsymbol{\psi}^{n,j,R}\|^2 + \|\boldsymbol{\psi}^{n,j,I}\|^2) \\
&\le C_7(\tau+\eta)^2\sum_{m=1}^{M+1}(\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2), \qquad (40)
\end{aligned}
$$

where $C_6 = JC_3\max\{1, 2C_4\}$ and $C_7 = 2JC_3\max\{(C_0)^2 + (C_2)^2\}\max\{1, 2C_4\}$. Finally we obtain (26) and (27). □

Now, we are ready to prove Theorem 1 in terms of the two lemmas above.

Proof of Theorem 1. First we prove (i). By (24)–(27) and Taylor's formula we have

$$
\begin{aligned}
&E(\mathbf{W}^{n+1}) - E(\mathbf{W}^n) \\
&= \sum_{j=1}^{J}(g_{jR}(S^{n+1,j,R}) - g_{jR}(S^{n,j,R}) + g_{jI}(S^{n+1,j,I}) - g_{jI}(S^{n,j,I})) \\
&= \sum_{j=1}^{J}\big(g'_{jR}(S^{n,j,R})(S^{n+1,j,R} - S^{n,j,R}) + g'_{jI}(S^{n,j,I})(S^{n+1,j,I} - S^{n,j,I}) \\
&\qquad + \frac{1}{2}g''_{jR}(t_1^{n,j})(S^{n+1,j,R} - S^{n,j,R})^2 + \frac{1}{2}g''_{jI}(t_2^{n,j})(S^{n+1,j,I} - S^{n,j,I})^2\big)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{j=1}^{J} \Big(g'_{jR}(S^{n,j,R}) \Big((\mathbf{H}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,R} - (\mathbf{H}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,I} \Big) \\
&\qquad + g'_{jI}(S^{n,j,I}) \Big((\mathbf{H}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,I} + (\mathbf{H}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,R} \Big) \\
&\qquad + g'_{jR}(S^{n,j,R}) \Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \mathbf{w}_{M+1}^{n,R} - (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \mathbf{w}_{M+1}^{n,I} \Big) \\
&\qquad + g'_{jI}(S^{n,j,I}) \Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \mathbf{w}_{M+1}^{n,I} + (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \mathbf{w}_{M+1}^{n,R} \Big) \\
&\qquad + g'_{jR}(S^{n,j,R}) \Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,R} - (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,I} \Big) \\
&\qquad + g'_{jI}(S^{n,j,I}) \Big((\boldsymbol{\psi}^{n,j,R})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,I} + (\boldsymbol{\psi}^{n,j,I})^{\mathbf{T}} \Delta \mathbf{w}_{M+1}^{n+1,R} \Big) \\
&\qquad + \frac{1}{2} g''_{jR}(t_1^{n,j})(S^{n+1,j,R} - S^{n,j,R})^2 + \frac{1}{2}\, g''_{jI}(t_2^{n,j})(S^{n+1,j,I} - S^{n,j,I})^2 \Big) \\
&\le (\tau - \eta + (C_6 + C_7)(\tau + \eta)^2)\big(\|\mathbf{p}_{M+1}^{n,R}\|^2 + \|\mathbf{p}_{M+1}^{n,I}\|^2\big) \\
&\qquad + (\tau - \eta + (C_5 + C_6 + C_7)(\tau + \eta)^2) \sum_{m=1}^{M} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2) \qquad (41)
\end{aligned}
$$

where $t_1^{n,j} \in \mathbb{R}$ is on the segment between $S^{n+1,j,R}$ and $S^{n,j,R}$, and $t_2^{n,j} \in \mathbb{R}$ is on the segment between $S^{n+1,j,I}$ and $S^{n,j,I}$.

Obviously, by choosing $C^{\star} = C_5 + C_6 + C_7$ and the learning rate η to satisfy

$$0 < \eta < \frac{1 - s}{C^{\star}(1 + s)^2}, \qquad (42)$$

we have

$$\tau - \eta + (C_6 + C_7)(\tau + \eta)^2 \le 0,\ \ \tau - \eta + (C_5 + C_6 + C_7)(\tau + \eta)^2 \le 0.$$

Then there holds

$$E(\mathbf{W}^{n+1}) \le E(\mathbf{W}^n),\ n = 0, 1, 2, \cdots \qquad (43)$$

Conclusion (ii) immediately results from Conclusion (i) since $E(\mathbf{W}^n) \ge 0$.

Next, we prove (iii). In the following we suppose (42) is valid. Let $\alpha = -(\tau - \eta + (C_6 + C_7)(\tau + \eta)^2)$, $\beta = -(\tau - \eta + (C_5 + C_6 + C_7)(\tau + \eta)^2)$, then $\alpha \ge 0$, $\beta \ge 0$. According to (41), we have

$$
\begin{aligned}
E(\mathbf{W}^{n+1}) &\le E(\mathbf{W}^n) - \alpha(\|\mathbf{p}_{M+1}^{n,R}\|^2 + \|\mathbf{p}_{M+1}^{n,I}\|^2) - \beta \sum_{m=1}^{M} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2) \\
&\le \cdots \le E(\mathbf{W}^0) - \sum_{k=0}^{n} \Big(\alpha(\|\mathbf{p}_{M+1}^{k,R}\|^2 + \|\mathbf{p}_{M+1}^{k,I}\|^2) + \beta \sum_{m=1}^{M} (\|\mathbf{p}_m^{k,R}\|^2 + \|\mathbf{p}_m^{k,I}\|^2) \Big). \qquad (44)
\end{aligned}
$$

Since $E(\mathbf{W}^{n+1}) \ge 0$, there holds

$$\sum_{k=0}^{n} \Big(\alpha(\|\mathbf{p}_{M+1}^{k,R}\|^2 + \|\mathbf{p}_{M+1}^{k,I}\|^2) + \beta \sum_{m=1}^{M} (\|\mathbf{p}_m^{k,R}\|^2 + \|\mathbf{p}_m^{k,I}\|^2) \Big) \le E(\mathbf{W}^0).$$

Let $n \to \infty$, then

$$\sum_{k=0}^{\infty} \big(\alpha(\|\mathbf{p}_{M+1}^{k,R}\|^2 + \|\mathbf{p}_{M+1}^{k,I}\|^2) + \beta \sum_{m=1}^{M} (\|\mathbf{p}_m^{k,R}\|^2 + \|\mathbf{p}_m^{k,I}\|^2)\big) \le E(\mathbf{W}^0) < \infty.$$

Hence, there holds

$$\lim_{n\to\infty} (\|\frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}_m^R}\|^2 + \|\frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}_m^I}\|^2) = \lim_{n\to\infty} (\|\mathbf{p}_m^{n,R}\|^2 + \|\mathbf{p}_m^{n,I}\|^2) = 0,$$

which implies

$$\lim_{n\to\infty} \|\frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}_m^R}\| = \lim_{n\to\infty} \|\frac{\partial E(\mathbf{W}^n)}{\partial \mathbf{w}_m^I}\| = 0, \quad 1 \le m \le M+1. \tag{45}$$

Finally, we prove (iv). We use (14), (15), (28), (29), and (45) to obtain

$$\lim_{n\to\infty} \|\mathbf{w}_m^{n+1,R} - \mathbf{w}_m^{n,R}\| = 0, \lim_{n\to\infty} \|\mathbf{w}_m^{n+1,I} - \mathbf{w}_m^{n,I}\| = 0, \ m = 1, \cdots, M+1. \tag{46}$$

Write $\mathbf{v} = ((\mathbf{w}_1^R)^\mathbf{T}, \cdots, (\mathbf{w}_{M+1}^R)^\mathbf{T}, (\mathbf{w}_1^I)^\mathbf{T}, \cdots, (\mathbf{w}_{M+1}^I)^\mathbf{T})^\mathbf{T}$, then $E(\mathbf{W})$ can be viewed as a function of $\mathbf{v}$, and denoted as $\mathcal{E}(\mathbf{v})$:

$$E(\mathbf{W}) \equiv \mathcal{E}(\mathbf{v}).$$

Obviously, $\mathcal{E}(\mathbf{v})$ is a continuously differentiable real-valued function and

$$\frac{\partial \mathcal{E}(\mathbf{v})}{\partial \mathbf{w}_m^R} \equiv \frac{\partial E(\mathbf{W})}{\partial \mathbf{w}_m^R}, \frac{\partial \mathcal{E}(\mathbf{v})}{\partial \mathbf{w}_m^I} \equiv \frac{\partial E(\mathbf{W})}{\partial \mathbf{w}_m^I}, \ m = 1, \cdots, M+1.$$

Let $\mathbf{v}^n = ((\mathbf{w}_1^{n,R})^\mathbf{T}, \cdots, (\mathbf{w}_{M+1}^{n,R})^\mathbf{T}, (\mathbf{w}_1^{n,I})^\mathbf{T}, \cdots, (\mathbf{w}_{M+1}^{n,I})^\mathbf{T})^\mathbf{T}$, then by (45) we have

$$\lim_{n\to\infty} \|\frac{\partial \mathcal{E}(\mathbf{v}^n)}{\partial \mathbf{w}_m^R}\| = \lim_{n\to\infty} \|\frac{\partial \mathcal{E}(\mathbf{v}^n)}{\partial \mathbf{w}_m^I}\| = 0, \quad m = 1, \cdots, M+1. \tag{47}$$

From Assumption (A3), (46), (47) and Lemma 1 we know that there is a $\mathbf{v}^\star$ satisfying $\lim_{n\to\infty} \mathbf{v}^n = \mathbf{v}^\star$. By considering the relationship between $\mathbf{v}^n$ and $\mathbf{W}^n$, we immediately get the desired result. We thus complete the proof. □

References

[1] Mandic D. P., Goh S. L.: Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models. NewYork: Wiley, 2009.

[2] Hirose A.: Complex-Valued Neural Networks. New York: Springer-Verlag New York Inc., 2006.

[3] Sanjay S. P. R., William W. H.: Complex-valued neural networks for nonlinear complex principal component analysis. Neural Networks, **18**, 2005, pp. 61-69.

[4] George G. M., Koutsougeras C.: Complex Domain Backpropagation. IEEE Trans. Circuits and Systems – II: Analog and Digital Signal Processing, **39**, 5, 1992, pp. 330-334.

[5] Benvenuto T., Piazza F.: On the complex backpropagation algorithm. IEEE Trans. Signal Processing, **40**, 4, 1992, pp. 967-969.

[6] Kim T., Adali T.: Fully complex backpropagation for constant envelop signal processing. In: Proc. IEEE Workshop Neural Network Signal Process., Sydney, Australia, 2000, pp. 231-240.

[7] Kim T., Adali T.: Nonlinear satellite channel equalization using fully complex feed-forward neural networks. In: Proc. IEEE Workshop Nonlinear Signal Image Process., Baltimore, MD, 2001, pp. 141-150.

[8] Yang S. S., Siu S., Ho C. L.: Analysis of the initial values in split-complex backpropagation algorithm. IEEE Trans. Neural Networks, **19**, 9, 2008, pp. 1564-1573.

[9] Wu W., Feng G. R., Li Z. X., Xu, Y. S.: Deterministic convergence of an online gradient method for BP neural networks. IEEE Trans. Neural Networks, **16**, 3, 2005, pp. 533-540.

[10] White H.: Some asymptotic results for learning in single hidden-layer feedforward network models. J. Amer. Statist. Assoc., **84**, 1989, pp. 1003-1013.

[11] Zhang H., Wu W., Liu F., Yao M.: Boundedness and convergence of online gadient method with penalty for feedforward neural networks. IEEE Trans. Neural Networks, **20**, 6, 2009, pp. 1050-1054.

[12] Mandic D. P., Chambers J. A.: Recurrent neural networks for prediction: learning algorithms, architectures and stability. NewYork: Wiley, 2001.

[13] Goh S. L., Mandic D. P.: A complex-valued RTRL algorithm for recurrent neural networks. Neural Computation. **16**, 12, 2004, pp. 2699-2713.

[14] Xu D. P., Zhang H. S., Liu L. J.: Convergence analysis of three classes of split-complex gradient algorithms for complex-valued recurrent neural networks. Neural Computation, **22**, 10, 2010, pp. 2655-2677.

[15] Zhang H. S., Xu D. P., Wang, Z. P.: Convergence of an online split-complex gradient algorithm for complex-valued neural networks. Discrete Dyn. Nat. Soc., 2010, pp. 1-27, DOI: 10.1155/2010/829692.

[16] Polyak B. T.: Introduction to Optimization, Optimization Software. New York: Inc. Publications Division, 1987.

[17] Shi Z. J., Guo J.: Convergence of memory gradient method. International Journal of Computer Mathematics, **85**, 2008, pp. 1039-1053.

[18] Qian N.: On the momentum term in gradient descent learning algorithms. Neural Networks, **12**, 1999, pp. 145-151.

[19] Wu W., Zhang N. M., Li Z. X., Li L., Liu Y.: Convergence of gradient method with momentum for back-propagation neural networks. Journal of Computational Mathematics, **26**, 2008, pp. 613-623.

[20] Ujang B. C., Took C. C., Mandic D. P.: Split quaternion nonlinear adaptive filtering. Neural Networks, **23**, 2010, pp. 426-434.

[21] Ortega J. M., Rheinboldt W. C.: Iterative Solution of Nonlinear Equations in Several Variables. New York: Academic Press, 1970.

[22] Nitta T.: Orthogonality of decision boundaries in complex-valued neural networks. Neural Computation, **16**, 2003, pp. 73-97.

[23] Savitha R., Suresh S., Sundararajan N.: A fully complex-valued radial basis function, International Journal of Neural Systems, **19**, 4, 2009, pp. 253-267.